科学结构地图的领域群自动识别研究*

王小梅[1]　邓启平[1,2]
[1](中国科学院文献情报中心　北京 100190)
[2](中国科学院大学　北京 100049)

摘要: 【目的】探索科学结构地图中研究领域群的自动识别方法, 快速勾勒科学结构全貌, 增加时效性。【方法】利用特征词测度研究领域的主题相似性, 同时考虑研究领域的相对位置关系, 将位置相邻、主题相似的研究领域划为领域群。设计有效性评价指标对比不同方法的最优参数组合, 推荐最优方法。【结果】该方法能有效地识别出不同时期科学结构地图的领域群。【局限】方法的有效性是基于"科学结构地图"数据的实验结果得到, 参数组合是否适用于其他数据还有待进一步验证。【结论】为科学结构地图领域群的自动识别提供了有效方法。

关键词: 科学结构　研究领域　领域群　自动识别
分类号: TP393　G35

1 引　言

科学知识结构体系一直是人们研究的热点, 科学结构地图通过可视化技术, 以直观形象的图谱形式展现高度抽象的科学, 特别是自然科学基础研究的宏观结构, 揭示了科学热点前沿间的关联关系与发展进程, 它可以帮助人们快速、全面和形象地把握科学总体态势。早在 1965 年, De Solla Price 就推测文献数据库中隐含着科学结构体系, 并提出通过整合期刊之间的引用关系可以揭示学科结构, 甚至更加细致地描绘研究方向[1]。随后, Carpenter 等通过对 SCI 数据库中的期刊进行聚类分析描绘出学科子领域[2], Small 等利用计算机技术对高被引科学文献聚类识别出整体的专业结构及之间的关联关系, 开启一个自动探测科学结构的新道路[3]。近年来, 随着文献计量方法的发展, 科学结构地图的研究也不断发展。

科学结构地图研究的一个重要问题是如何解读。目前绘制科学结构地图分析单元主要有引文、主题词、作者、期刊等, 其中同被引分析能更好地理解科学发现在某个专题发展中的作用[4]、更详细地描绘学科结构特征。实际应用中, 直接通过同被引聚类得到的以研究领域为基本单元的科学结构地图可读性不高, 因为将上百个研究领域放在一起, 读者很难直观地获取有用信息, 所以研究者通常会对每个研究领域进行内容分析, 并根据研究内容的相似性将众多的研究领域划分到不同的领域群(研究大类)中并对其命名, 绘制出整个科学结构的概貌, 反映不同研究大类间的关联关系。因此领域群的识别在科学结构地图研究中有着重要意义。然而, 随着科学的发展, 新兴研究领域不断出现和现有研究领域的消亡, 使科学结构地图也在发展变化中, 如何快速有效地绘制科学结构地图, 并识别其领域群是把握科学发展变化的关键和难点。已有的研究主要是采用人工判读方法, 本研究旨在探索一种自动方法识别科学结构地图中的领域群。

2 相关研究

2.1 科学结构地图的绘制步骤

(1) 分析单元的选择, 比如引文、关键词、作者、期刊等, 每种单元都有对应的共现分析, 如共被引、共词、共作者及期刊引用等。

(2) 确定这些分析单元之间的关联。

(3) 可视化显示, 把分析单元及它们之间的关联

通讯作者: 王小梅, ORCID: 0000-0002-9895-1511, E-mail: wangxm@mail.las.ac.cn。

*本文系国家自然科学基金项目"科学结构特征及其演化动力学分析方法与应用研究"(项目编号: 71173211)的研究成果之一。

关系在低维空间(通常是二维)里显示。

本文涉及的科学结构地图由作者所在研究团队自主开发，具体绘制方法参见文献[5-8]。其分析单元是引文，通过对汤森路透集团基本科学指标库(Essential Science Indicators, ESI)中高被引论文的同被引聚类，形成包含若干研究论文的“研究领域”；采用重力模型算法确定各个研究领域在二维空间中的布局位置，第一期科学结构布局计算时初始位置固定，后几期布局采用与前一期平行映射的方法，保证布局位置的稳定和可对比。研究领域间的相对位置关系反映了它们的关联程度，距离越近相关性越强。

2.2 领域群识别方法

通过上述流程绘制的科学结构地图由于研究领域数量比较多，无法在图中标识每个研究领域的名称，因此其提供的直接信息有限，为了增加其可读性，研究者通常需要将研究领域划分到不同的领域群中。领域群也称为研究大类或类学科结构，是更高层次的科学结构，识别科学结构地图中的领域群能绘制出整个科学结构的概貌，便于研究者按研究大类观察科研态势。在基于引文聚类生成的科学结构地图中，领域群识别通常采用人工判读划分领域群，自动识别方法的研究还处于探索阶段。

(1) 人工判读识别领域群

人工判读是常用的领域群识别方法，研究者将引文聚类得到的论文列表以及从中抽取的关键词信息提供给相应的领域专家进行判读，领域专家根据提供的关键词和论文信息对每个研究领域进行命名并将其划分到不同的研究大类。根据判读结果将科学结构地图中属于同一研究大类下的研究领域借助画图软件画入不规则区域中生成领域群。在科学结构地图系列专著《科学结构地图 2009》[6]、《科学结构地图 2012》[7]、《科学结构地图 2015》[8]以及日本科学技术政策研究所(NISTEP)关于科学结构演化的类似研究[9]中均采用人工判读的方法识别领域群。人工判读识别领域群的结果最为准确，但其工作流程繁琐，对领域专家的要求较高，会延迟科学结构地图的发布时间，因此亟需一种有效的自动识别方法来代替人工判读。

(2) 自动识别领域群

有关领域群自动识别方法的研究还很少，笔者所在研究团队在《科学结构地图 2015》[8]中在原有二次聚类基础上，尝试构建研究领域间的引用关系，利用研究领域之上的三次聚类自动识别领域群，但效果并不理想，分析认为该方法存在以下不足:

①三次聚类时，领域间的引用关系是领域中论文引用关系的合集，层次太高，这种引用关系就会有放大、失真，因此会影响其聚类的准确性;

②研究领域在科学结构地图中的相对位置反映了它们之间的关联程度，位置关系是识别领域群的重要条件，该方法没有考虑位置关系，识别结果的准确度偏低。

为了提升自动识别领域群的准确性，有研究者尝试利用研究领域间的相对位置关系识别领域群。Boyack 等定义类学科结构，利用一个半自动方法将研究领域划分到类学科结构组群中。其方法是将科学结构地图划分为网格，选取一系列特定的网格作为学科种子，以学科种子为中心，网格外接圆重叠部分包含的文献数量作为网格的连接机制，将其相邻的、共有文献最多的网格或网格群连接到一个组群中，重复该步骤直到所有的网格被连接到一个类学科结构中[10]，该方法虽然考虑了研究领域间的位置关系，但其过于依赖选取的学科种子数量，不同的学科种子数量下识别出的领域群差别很大。NISTEP 在 2014 年发表的《Science Map 2010 & 2012》中提出一种利用研究领域位置关系并结合主题相似性识别领域群的方法。研究同样将科学地图划为网格，并按照包含的论文数量对网格排序，依次计算网格与一定范围内其他网格具有的相同特征词个数，如果该值大于阈值则认为属于同一个候补领域群，重复上述步骤直到所有的网格都划分到候补领域群中，最后按照一定规则对候补领域群进行删除、合并得到最终领域群划分结果[11]。NISTEP 提出的方法结合了研究领域间的位置关系和主题相似性，能较准确地识别科学结构地图中的领域群，但笔者在试验中发现，该方法对科学结构地图中研究领域密集的区域区分度较低，且识别结果对参数敏感，由于该方法涉及多个参数，实际应用较为困难。

3 研究方法

研究领域在科学结构地图中的相对位置是通过布局算法得到，距离越近的研究领域引用关联性越强，更有可能属于同一领域群，因此位置关系是领域群识别中必不可少的因素。而利用位置关系反映的是同被引关系的远近，如果进一步借助文本分析判别

主题相似性，将提高划分领域群的准确性。针对上述调研方法的不足，本研究借鉴 NISTEP 在 2014 年提出的领域群识别方法，尝试两种主题相似性测度方法，改进候补领域群识别方法，意在提升研究领域密集区域的区分度与精确度，并从 F 值、领域群识别数量、领域群重叠情况等多个维度测评方法的有效性，通过对比分析找出最有效的识别方法及最优参数组合。

3.1 方法介绍

(1) 关联相似性测度

根据布局原理，位置距离越近的研究领域关联性越强。本研究根据研究领域的坐标范围将科学结构地图划分为网格，同一网格下的研究领域默认属于同一领域群。利用主题相似性建立网格间的连接机制，将指定范围内研究内容相似的网格连接成为领域群。划分网格的原则需要注意网格内研究领域的数量，网格稀疏研究领域数量较多，难以区分临近的领域群；网格密集研究领域数量较少，不能有效建立网格间的连接机制，识别出的领域群规模偏小，数量增多。

(2) 主题相似性测度

为了测度网格间研究内容的相似性，本研究采用了两种基于文本相似度的方法来测度网格的主题相似性。

①基于特征词数量测度主题相似性，利用 Alchemy API[12] 接口从论文的题目和摘要中抽取描述研究领域主题的特征词，对于任意需要测度相似性的两个网格，分别从包含的研究领域中选取出现频次最高的 n 个特征词，统计共同特征词数量作为二者之间的相似度，当相似度高于设定的阈值时则认为二者属于同一领域群；

②基于特征向量测度主题相似性，从研究领域的论文题目和摘要中提取出全部单词，去除停用词后转为特征向量，用特征向量夹角余弦值表示研究领域的主题相似度，当两个网格间研究领域的相似度的平均值大于设定阈值时认为二者主题相似，属于同一领域群。

(3) 候补领域群识别方法

本文提出一种动态的领域群识别方法，与上述基于位置的关联相似性和主题相似性结合改进候补领域群的识别精度。

①基于特征词的静态识别方法

NISTEP 提出的领域识别方法是以候补网格为中心，分别计算周围网格与候补网格的相似度，同时将主题相似的网格划为一个候补领域群，这种方法称为静态识别法。

原理如图 1 所示，从论文的题目和摘要中抽取出现频次最高的 60 个描述研究领域主题的特征词，根据包含的论文数量将所有网格降序排列作为候补网格，依次以每个候补网格为中心，在指定范围(领域群规模参数)内遍历网格，计算与候补网格主题相似的网格(共同特征词数大于阈值)，用椭圆将这些网格连接起来视为一个候补领域群，同时将它们从候补网格中删除，重复该步骤直到所有的网格都划入领域群中。

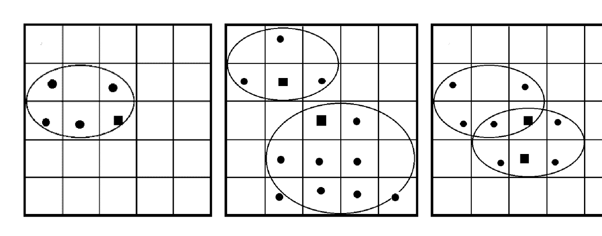


图 1 基于特征词相似的领域群识别

(注：图中黑色方块所在网格为候补网格，圆圈表示该网格与候补网格主题相似。)

②基于特征词的动态识别方法

上述方法用于测度主题相似的特征词集合是固定不变的，该方法对网格间的主题相似性程度没有区分，大于阈值的网格相似性程度被视为一致，同时划入一个领域群中。而科学结构地图中研究领域的分布并不均匀，一些研究领域密集的区域可能包含多个领域群，并且主题相似性程度差异较大，即各个领域群内相似性阈值差异较大。而稀疏区域的网格包含的共同特征词数相对较少，基于特征词的静态识别方法很难找到一个阈值将密集区域的领域群区分开的同时识别出稀疏区域的领域群。

为了解决上述不足，研究提出动态提取特征词的方法，其核心思想是候补领域群的特征词是动态变化的，在识别过程中将相似性最高的网格逐步划入候补领域群中，每一步迭代后重新抽取领域群的特征词，用动态的特征词测度主题的相似性。这种动态的过程可以更好地识别研究领域密集区域包含的领域群。如图 2 所示，将候补网格 A 周围的网格分为多层，用以限制领域群的规模。研究领域的距离越近，属于同一领域群的可能性越大，因此设定从第一层网格开始识别候补领域群。以候补网格 A 为中心，根据主题相似性测度方法将第一层网格中与网格 A 主题最相似的网格划入候补领域 A 中，图中网格 B6 与网格 A 主题相似度最高，将其划入候补领域群 A 中；重新抽取领域群 A 的特征词，测度剩余网格与领域群 A 的主题相似性，接着将 B4 划入候补领域群 A 中；重复该步骤直到第一层所有高于阈值的网格都划入候补领域群 A 中，并将划入候补领域群的网格从候补网格中删除。以同样的方法处理其他指定层数内的网格，最后得到以网格 A 为中心的候补领域群。

③基于特征向量的动态识别方法

本方法中，两个研究领域采用特征向量余弦夹角计算

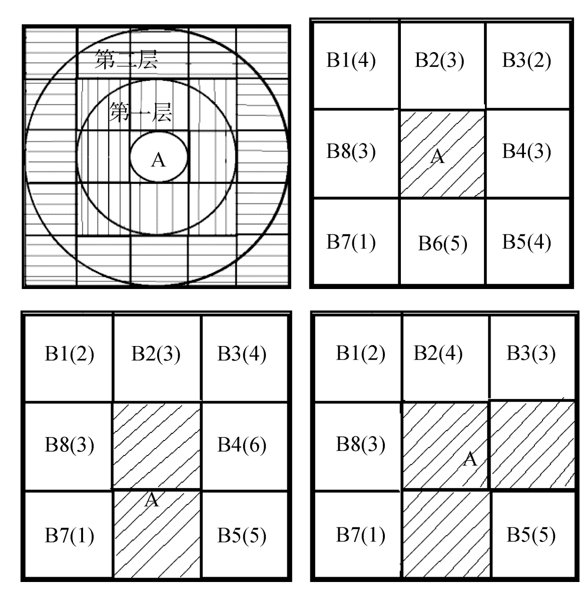


图 2　基于动态特征词识别领域群

其相似度，当两网格之间的研究领域相似度的平均值大于设定阈值时则认为两个网格主题相似。以候补网格为中心，将相似度最大的网格与候补相连，形成一个新的候补网格。如此迭代，将相似度最大的网格依次划入候补网格中形成领域群。

3.2　具体流程

(1) 将科学结构地图进行网格分割，通常划为 20*20 或 30*30 的网格，可根据研究领域的分布情况进行调整。

(2) 计算包含每个网格的密度(论文数)，根据密度进行降序排列作为候补网格。

(3) 按照候补领域群的识别方法，依次以候补网格为中心识别候补领域群，直到所有的候补网格都划入领域群中。一个领域群应该包含至少两个研究领域，因此将只包含一个研究领域的候补领域群删除。

(4) 求出候补领域群中研究领域在 X、Y 轴上的最大值和最小值以及中心点，以中心点坐标为中心，$(X_{max}-X_{min})$为 X 轴方向的长度，$(Y_{max}-Y_{min})$为 Y 轴方向的长度，绘制椭圆标识该领域群。

(5) 重叠领域群删除：由于一个网格可以属于多个领域群，候补领域群存在重叠的情况，因此需要删除被其他领域群覆盖的领域群：

①删除完全包含于其他领域群中的领域群；

②在椭圆等式为 $X^2/A^2+Y^2/B^2=1$ 的情况下，当另一椭圆中心点 x1、y1，满足条件 $x1^2/A^2+y1^2/B^2<0.5$ 时，删除面积较小的领域群；

③重新细化网格，如果一个领域群有超过 80%的网格包含于其他领域群，则删除该领域群。

(6) 重叠区域大的领域群合并：经过步骤(5)删除之后仍存在一些交叉重叠的领域群，当两个重叠领域群的重叠相似度大于合并阈值时，即合并两个领域群。基于特征词方法的合并阈值为 30 个共同特征词；基于特征向量方法的合并阈值为 0.15。其他小于阈值的重叠情况是允许存在的，因为这些重叠现象反映了研究内容的交叉性。

3.3　有效性测评

领域群的识别实质是将主题相似的研究领域划分在一起，本文的聚类是模糊聚类，即一个研究领域可以属于多个大类。因此本文利用修改的聚类分析效果评价指标 F 值[13]验证领域群识别结果的有效性。在实际应用中，希望自动识别的结果尽可能接近人工标识结果，反映科学结构中的主体结构，因此将领域群自动识别的结果与人工判读的结果进行比较验证。对于自动识别领域群不能对应人工标识结果的现象，通常是 1-2 个，是一些边缘上或人工觉得层次不够，不予以标注的聚类，但其应该也是有道理和有一定价值的，在本文的有效性评测中不参与计算。对每个人工标注的领域群 P_j，假设在自动识别结果中存在一个与之对应的领域群 A_i，这个对应关系未知。为了发现 A_i，遍历所有聚类结果，分别计算准确率、召回率和 F 值，从中挑选最优 F 值及其对应的领域群。进一步对所有领域群的 F 值作加权平均，得到整个识别结果的 F 值。对于人工标注的领域群 P_j, A_i 的准确率、召回率和 F 值为：

$$P(P_j,A_i)=\frac{|P_j\cap A_i|}{A_j}$$

$$R(P_j,A_i)=\frac{|P_j\cap A_i|}{P_j}$$

$$F(P_j,A_i)=\frac{2P(P_j,A_i)\cdot R(P_j,A_i)}{P(P_j,A_i)+R(P_j,A_i)}$$

P_j 的 F 值为：

$$F(P_j)=\max_{1\leqslant i\leqslant m}F(P_j,A_i)$$

整个结果 F 值的计算方法如公式(1)所示：

$$F=\sum_{j=1}^{s}w_j\cdot F(P_j),\ \ w_j=\frac{|P_j|}{\sum_{i=1}^{s}|P_i|}=\frac{P_j}{n} \tag{1}$$

原始的 F 值评价适用于每个人工标注领域群对应一个最优的自动识别领域群的情况，但实际上自动识

别领域群的数量大于人工识别领域群的数量，因此部分自动识别出的领域群在科学结构层级上低于人工标注领域群，存在多个自动识别领域群对应同一个人工标注领域群的情况。因此评价时将70%的研究领域包含于同一人工标注领域群的自动领域群进行合并，再计算 F 值。由于某些研究领域没有对应的领域群，本研究在F值的基础上进一步考虑了不同方法的区分度，用修正后的 F_{di} 表示方法的有效性。

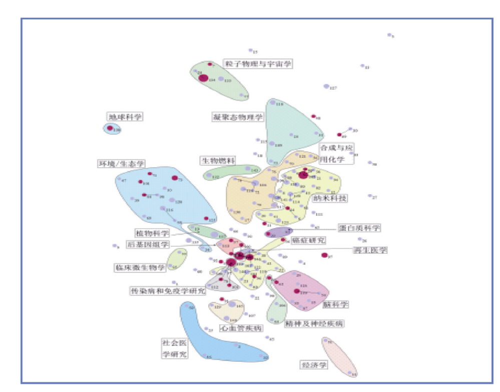
$$F_{di}=\frac{\text{总研究领域个数}-\text{未识别出领域群中研究领域个数}}{\text{总研究领域个数}}\times F \quad (2)$$

4 方法对比分析

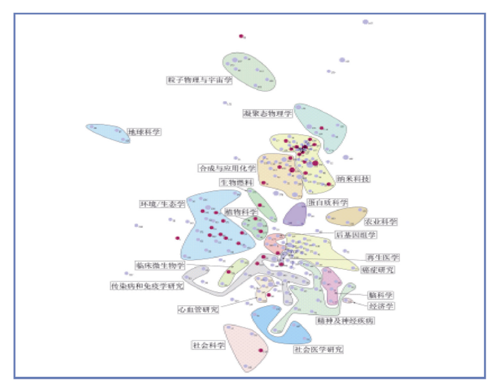
本文基于《科学结构地图 2015》[8]中的科学结构地图数据进行方法对比分析，按照上文描述的方法与流程自编程序，绘制出不同时期与不同参数下的领域群效果图，利用修正的 F 值指标比较分析改进后的方法与原方法的有效性以及不同参数值对方法有效性的影响，找出方法中最优的参数组合。

4.1 数 据

研究选取科学结构地图 2006~2011 和科学结构地图 2008~2013 两个时期的数据进行实验分析，如图 3 所示。图 3 中每一个圆代表一个研究领域，圆的大小与研究领域包含的论文数量成正比，圆旁边的数字代表研究领域的 ID 号。科学结构地图 2006~2011 含有 149 个研究领域，科学结构地图 2008~2013 含有 212 个研究领域，将其人工划分为 10 个大类(包含 2 个及以上研究领域的领域群)，每个大类为一个不规则线区域，用不同颜色区分。大类名称如表 1 所示。



(a) 2006~2011时期



(b) 2008~2013时期

图 3 科学结构地图

表 1 科学结构地图大类表

时期 \ 大类名称及领域数 \ ID	1	2	3	4	5	6	7	8	9	10
2006~2011	粒子物理与宇宙学	凝聚态物理学	纳米科技	合成与应用化学	环境/生态学	生物学	医学	经济学	工程科学	数学
	5	5	22	14	14	11	52	2	9	3
2008~2013	粒子物理与宇宙学	凝聚态物理学	纳米科技	合成与应用化学	环境/生态学	生物学	医学	社会科学	工程科学	农业科学
	7	6	35	18	25	18	77	4	8	3

4.2 有效性分析

研究从 F_{di} 值、领域群数量以及重叠情况等多个维度对三种方法有效性进行验证，包括不同方法、不同参数之间的对比，以找出最优方法中的最优参数组合。

结构地图大小为 360*480 像素，为了使研究领域适应画布大小，实验中对研究领域坐标做了平移处理。

可调参数主要包括网格数、相似度阈值和领域群规模，不同识别方法中各参数设定有所不同。本研究将网格划分为 20*20，30*30 两种；基于特征词方法共同特征词数量阈值选取 2 个、3 个、4 个词进行对比，基于特征向量方法的相似度阈值选取 0.07 和 0.12 两个值进行对比；领域群规模在静态方法中参考NSTEP方法用绝对距离表示，经过反复试验，将其设定为 70 像素效果最佳，在动态方法中用网格层数表示，是一个相

对距离，可根据网格的疏密进行调整，试验中将该值设定为 2。表 3 和表 4 是两个时期科学结构地图的领域群识别结果的整体 F_{di} 值。

表 3　科学结构地图 2006~2011 领域群识别结果的 F_{di} 值对比

阈值 \ F_{di} 值 \ 网格	动态特征词		静态特征词		特征向量	
	20*20	30*30	20*20	30*30	20*20	30*30
2	0.64	0.63	0.54	0.62	/	/
3(特征向量 0.07)	0.70	0.62	0.57	0.49	0.68	0.65
4(特征向量 0.12)	0.68	0.59	0.53	0.34	0.63	0.61

表 4　科学结构地图 2008~2013 领域群识别结果的 F_{di} 值对比

阈值 \ F_{di} 值 \ 网格	动态特征词		静态特征词		特征向量	
	20*20	30*30	20*20	30*30	20*20	30*30
2	0.71	0.72	0.60	0.54	/	/
3(特征向量 0.07)	0.70	0.72	0.64	0.61	0.68	0.65
4(特征向量 0.12)	0.66	0.69	0.50	0.52	0.77	0.76

分析可以看出，在两个不同时期的科学结构地图下，研究提出的基于动态的识别方法整体上占优。其中基于动态特征词的识别结果的 F_{di} 值较高且变化范围较小，表明该方法不仅识别结果更为精准而且对参数的敏感度较低，识别结果更加稳定。对比两个时期，当科学结构地图中研究领域数量较多时(2008~2013 时期)，该方法在 30*30 的网格下识别的效果更好，此时随着共同特征词阈值的增加 F_{di} 值减小；而当研究领域数量较少时(2006~2011 时期)，20*20 网格下的 F_{di} 值更高一些，共同特征词阈值设为 3 时效果最好。

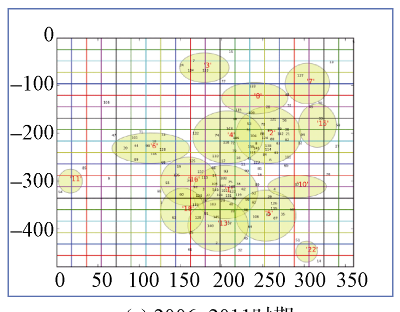
从领域群数量来看，总体趋势是随着网格密度和相似度阈值的增加，识别出的领域群数增多，规模变小。同一时期不同方法和参数之间的领域群数量变化范围不大，2008~2013 时期识别出的领域群数量在 13-16 个之间，由于 2006~2011 时期研究领域数量较少且分布稀疏，识别出的领域群数量略多于前者，在 14-19 个之间。相同参数下，基于动态特征词的方法识别出的领域群数量多于基于静态特征词方法，因为前者能更加稳定地将凝聚态物理学、纳米科技以及合成应用化学等关联性较强的领域区分开。

领域群的重叠情况与领域群数量变化类似，随着网格密度和相似度阈值的增加，重叠部分增加。对比不同方法，基于特征向量方法重叠度最低，其次是基于静态特征词方法，基于动态特征词方法的重叠度最高，其重叠多出现在关联性较强和交叉学科领域之间，分析认为这些重叠是合理的。

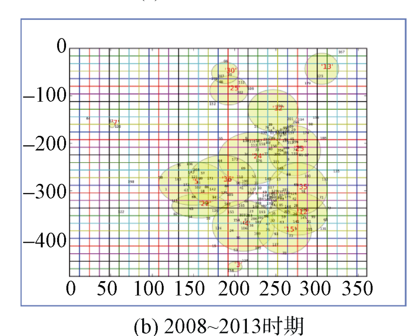
综合看来，基于特征词的静态识别方法在研究领域密集区域网格间的主题相似性程度差异较大，简单地将高于阈值的网格划入领域群中导致识别结果的区分度较低，识别出的领域群数量偏少，重叠度较低，F_{di} 值最小，且该方法对参数更加敏感，尤其是将网格划分为 30*30 时，随着相似度阈值的增加，领域群数量会快速增加，识别效果显著降低。基于动态特征词方法能有效识别出关联性较强的学科领域，识别出的领域群数量多于基于静态特征词方法，重叠度较高，F_{di} 值较大。相较于其他两种方法，基于特征向量方法的参数选取较为困难，虽然在 2008~2013 时期 F_{di} 值较高，但两时期科学结构地图的识别效果差异很大，说明该方法不稳定，对节点量小的网络区分效果不是很好。

4.3　结　果

可以看出，改进后的基于特征词的动态识别方法能较稳定、有效地识别科学结构地图的领域群，研究分别选取该方法最优参数组合下识别的两时期领域群效果图(见图 4)作为实验结果进行说明，其中椭圆表示自动识别出的领域群，椭圆中间的红色数字代表领域群的 ID 号，黑色数字代表研究领域的 ID 号，表 5 给出了人工标注的领域群和自动识别结果的对应关系及其 F 值。



(a) 2006~2011时期



(b) 2008~2013时期

图 4　领域群识别效果图(网格: 20*20，共同特征词: 3)

表 5　领域群 ID 对应关系

	人工标注 ID	1	2	3	4	5	6	7	8	9	10
自动识别 ID	2006~2011	3	8	2	4	6	16	1, 18, 13, 5	22	15	7
	$F(P_j)$	0.91	0.67	0.77	0.59	0.96	0.36	0.88	1.00	0.50	0.29
	2008~2013	25, 30	32	23	24	5, 7	36	15, 4, 35	3	13	35
	$F(P_j)$	0.88	0.38	0.81	0.55	0.85	0.44	0.88	0.67	0.18	0.13

研究重点关注了改进后的方法对粒子物理与宇宙学、凝聚态物理学、纳米科技、合成与应用化学、环境/生态学、生物学、医学等 7 个较大领域群的识别效果。从两时期科学结构地图的实验结果可以看出，该方法都能准确识别出 7 个领域群的位置，粒子物理与宇宙学、环境/生态学 2 个领域群，群内研究领域密度大，与其他领域群的关联程度弱，能被精确识别；凝聚态物理学、纳米科技、合成与应用化学 3 个领域群所在区域的研究领域十分密集，研究内容关联程度高，基于特征词的静态识别方法很难将其区分开，通常只能识别出 2 个领域群，改进的方法能有效地区分出 3 个领域群，但它们之间的重叠部分较多，规模较小的领域群准确率偏低；生物学的研究领域分布相对松散且与医学存在交叉，致使两时期科学结构地图中该领域群的 F 值都不高，但都识别出其研究领域集中分布的区域；医学是包含研究领域最多的交叉学科，改进后的方法将其所在区域划为多个领域群，研究认为这是合理的，因为从领域群规模来看，医学相较于其他领域群是更高层级的科学结构，其可以细分为多个子领域群。分析注意到，在科学结构地图 2008~2013 的识别结果中，工程科学的 F 值只有 0.18，是由于其包含的研究领域分布在地图的不同区域，导致不能有效识别。

5　结　语

本文探索了以研究领域为基本单元的科学结构地图的领域群自动识别方法，并通过实验对比分析找出方法中最优的参数组合。结合研究领域间的相对位置关系和主题相关性，将地图划分为网格，并利用特征词测度研究主题相似性，以此建立网格之间的连接机制，采用三种不同的领域群识别方法自动划分出领域群。利用聚类分析中基于人工标注簇的 F 值评测领域群自动识别方法的有效性，将不同方法的识别结果与人工判别的领域群进行对比分析。

对比不同方法的识别结果，发现研究提出的改进的动态识别方法相对准确地识别出科学结构地图中的领域群，尤其是动态特征词方法对研究领域密集的区域有较好的区分度，且对参数敏感度较低，比较稳定。基于特征向量方法虽然在某种情况下 F_{di} 值较高，但在不同数据集下识别结果不稳定，当没有人工判别的簇可供参考时，难以选择最优参数，且当数据量很大时计算复杂度高。应用时可以根据实际情况进行方法的选择或使用两种方法的结合来确定合适的参数。

研究中有关方法的结论均是以《科学结构地图 2015》[8]中的数据实验得到，后续进一步对其他数据集的科学结构地图进行验证。

参考文献：

[1] De Solla Price D J. Networks of Scientific Papers [J]. Science, 1965, 149(3683): 510-515.

[2] Carpenter M P, Narin F. Clustering of Scientific Journals [J]. Journal of the American Society for Information Science, 1973, 24(6): 425-436.

[3] Small H, Griffith B C. The Structure of Scientific Literatures I: Identifying and Graphing Specialties [J]. Science Studies, 1974, 4(1): 17-40.

[4] 陈超美. 科学前沿图谱知识可视化探索[M]. 北京：科学出版社, 2014. (Chen Chaomei. Mapping Scientific Frontiers: The Quest for Knowledge Visualization [M]. Beijing: Science Press, 2014.)

[5] 王小梅，韩涛，王俊，等. 基于同被引分析的科学结构图[J]. 科学观察, 2009, 4(4): 1-15. (Wang Xiaomei, Han Tao, Wang Jun, et al. Mapping Science Based on Co-citation Analysis [J]. Science Focus, 2009, 4(4): 1-15.)

[6] 潘教峰，张晓林，王小梅，等. 科学结构地图 2009[M]. 北京：科学出版社, 2010: 12-18. (Pan Jiaofeng, Zhang Xiaolin, Wang Xiaomei, et al. Mapping Science Structure 2009[M]. Beijing: Science Press, 2010: 12-18.)

[7] 潘教峰，张晓林，王小梅，等. 科学结构地图 2012[M]. 北京：科学出版社, 2013: 13-18. (Pan Jiaofeng, Zhang Xiaolin,

Wang Xiaomei, et al. Mapping Science Structure 2012[M]. Beijing: Science Press, 2013: 13-18.)
[8] 王小梅, 韩涛, 王俊, 等. 科学结构地图 2015[M].北京: 科学出版社, 2015: 10-34. (Wang Xiaomei, Han Tao, Wang Jun, et al. Mapping Science Structure 2015[M]. Beijing: Science Press, 2015: 10-34.)
[9] NISTEP. Science Map 2008[R/OL]. [2010-05-11]. http://data.nistep.go.jp/dspace/bitstream/11035/686/1/NISTEP-NR139-FullJ.pdf.
[10] Boyack K W, Klavans R. Creation of a Highly Detailed, Dynamic, Global Model and Map of Science[J]. Journal of the Association for Information Science and Technology, 2014, 65(4): 670-685.
[11] NISTEP. Science Map 2010 & 2012 [R/OL]. [2014-07-11]. http://data.nistep.go.jp/dspace/bitstream/11035/2933/4/NISTEP-NR159-FullJ.pdf.
[12] Alchemy API. AlchemyLanguage Features [EB/OL]. [2014-10-15]. http://www.alchemyapi.com/products/alchemylanguage.
[13] 周昭涛. 文本聚类分析效果评价及文本表示研究[D]. 北京: 中国科学院研究生院, 2005. (Zhou Zhaotao. Quality Evaluation of Text Clustering Result and Investing on Text Representation [D]. Beijing: The Graduate School of Chinese Academy of Sciences, 2005.)

作者贡献声明:

王小梅: 提出研究思路, 设计研究方案, 论文最终版本修订;
邓启平: 采集、清洗数据, 算法实现, 论文起草。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: wangxm@mail.las.ac.cn。

[1] 王小梅, 邓启平. RACoordinate2006_2011.csv. "科学结构地图 2006~2011"研究领域坐标.
[2] 王小梅, 邓启平. RASize2006_2011.csv. "科学结构地图 2006~2011"研究领域文献数量.
[3] 王小梅, 邓启平. RACategories2006_2011.csv. "科学结构地图 2006~2011"研究领域学科划分.
[4] 王小梅, 邓启平. FeatureWords2006_2011.json. "科学结构地图 2006~2011"研究领域特征词.
[5] 王小梅, 邓启平. RAGroups2006_2011.csv. "科学结构地图 2006~2011"研究群识别结果.
[6] 王小梅, 邓启平. RACoordinate2008_2013.csv. "科学结构地图 2008~2013"研究领域坐标.
[7] 王小梅, 邓启平. RASize2008_2013.csv. "科学结构地图 2008~2013"研究领域文献数量.
[8] 王小梅, 邓启平. RACategories2008_2013.csv. "科学结构地图 2008~2013"研究领域学科划分.
[9] 王小梅, 邓启平. FeatureWords2008_2013.json. "科学结构地图 2008~2013"研究领域特征词.
[10] 王小梅, 邓启平. RAGroups2008_2013.csv. "科学结构地图 2008~2013"研究群识别结果.

收稿日期: 2015-11-12
收修改稿日期: 2016-02-22

Auto-Identifying Research Area Groups in Science Map

Wang Xiaomei[1] Deng Qiping[1,2]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [**Objective**] This paper aims to establish an automatic method to identify research area groups and outline the science map quickly. [**Methods**] First, we used feature words to measure topic similarity, and then divided adjacent research areas with similar/related topics into groups. Second, we designed an effectiveness evaluation index to compare different optimal parameters combination. [**Results**] The proposed method could identify research area groups in science maps effectively. [**Limitations**] Our study was conducted with data from Mapping Science Structure 2015. More research is needed to investigate the proposed method's compatibility with other cases. [**Conclusions**] The proposed method could automatically identify research area groups in the science map.

Keywords: Science map Research area Research area groups Automatic detection